7. Regression Analysis

The term regression was first used by **Sir Francis Galton in 1877.** *Regression analysis is a form of predictive modeling technique which investigates the relationship between a* ***dependent*** *(target) and* ***independent variable*** *(s) (predictor).* It is concerned with the estimation of one variable for a given value of another variable on the basis of an average mathematical relationship between the two variables (or a number of variables). This technique is used for forecasting, time series modeling and finding **the causal effect relationship** between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

> ***According to M.M. Blair*** *"Regression is the measure of average relationship between two or more variables in term of the original unit of data.*

> ***According the Wallis and Roberts*** *"It is often more important to find out what relationship actually is in order to estimate or predict one variable (the dependent Variable) and the statistical technique appropriate to such called regression analysis.*

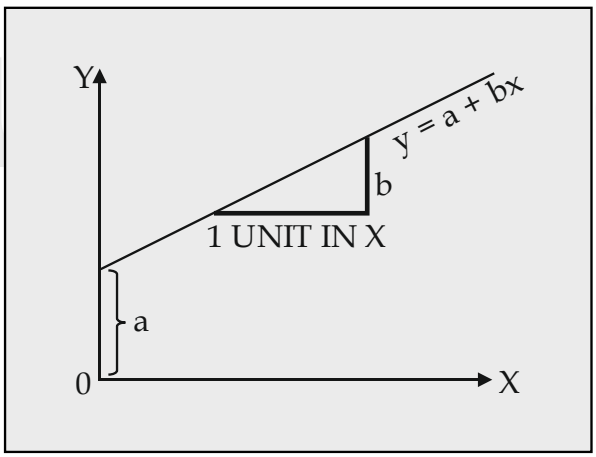
Thus regression is mathematical measure of the average relationship between a series of two or more variables in term of the original units of data under study. It is used to predict the value of one variable on the basis of the other.

If there are two variables say X and Y and if Y is influenced by X i.e. Y depends on X, then we get a simple linear regression. Here Y is known as dependent variable or explained variable and X is known as independent variable or explanatory variable.

In case of simple regression if Y depends on X, then the regression line of Y on X is given by:

$$Y = a + bX$$

Here 'a' and 'b' are two constants also known as regression parameters. 'b' is also known as regression coefficient of Y on X and is denoted by b_{yx}. The regression line is also known as 'line of best fit' and can be obtained by method of least square. The geometric presentation of regression line is as follows :



The two Normal Equations used to evaluate the values of 'a' and 'b' are:

$$\Sigma y = Na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Solving these two equations for 'b' and 'a', we will get the "least square" estimations of 'b' and 'a' as:

Focus Formula

$$b \text{ or } b_{yx} = \frac{Cov(x,y)}{S^2}$$
$$= \frac{r.S_x S_y}{S_x^2}$$
$$= \frac{r.S_y}{S_x}$$
$$a = \bar{y} - b\bar{x}$$

In case if X depends on Y, then the regression line of X on Y is given by:

$$X = a + bY$$

Here 'b' is the regression coefficient of X on Y and is denoted by b_{xy}. The two Normal Equations are as follows:

$$\sum x = Na + b\Sigma y$$
$$\sum xy = a\Sigma y + b\Sigma y^2$$

The values of 'b' and 'a' are as follows:

Focus Formula

$$b \text{ or } b_{yx} = \frac{Cov(x,y)}{S^2}$$
$$= \frac{r.S_x S_y}{S_y^2}$$
$$= \frac{r.S_x}{S_y}$$
$$a = \bar{x} - b\bar{y}$$

A single formula for estimation of b is as follows:

Focus Formula

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \text{ or } b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$\text{and} \quad b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2} \text{ or } b_{xt} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

The equation of regression lines can also be written as:

Y on X:

$$\boxed{Y - \bar{Y} = b_{yx}\left(X - \bar{X}\right)}$$

X on Y:

$$\boxed{X - \bar{X} = b_{xy}\left(Y - \bar{Y}\right)}$$

Properties of Regression Coefficient

1. *The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically :*

$$r = \sqrt{b_{xy} \times b_{yx}}$$

2. *If one of the regression coefficients is greater than unity, the other must be less than unity, since the value of the coefficient of correlation cannot exceed unity.*
3. *Both the regression coefficients will have the same sign, i.e., they will be either positive or negative.*
4. *The coefficient of correlation will have the same sign as that of regression coefficients, i.e., if regression coefficients have a negative sign, r will also have negative sign and if the regression coefficients have a positive sign, r would also be positive,*
5. *The average value of the two regression coefficients would be greater than the value of coefficient of correlation. In symbols* $(b_{xy}+b_{yx})/2>r$.
6. *Regression coefficients are independent of change of origin but not scale.*

Difference Between Regressions and Correlation

- Correlation is a statistical measure which determines co-relationship or association of two variables. Regression describes how an independent variable is numerically related to the dependent variable.
- Correlation represents linear relationship between two variables. Regression estimates one variable on the basis of another variable.
- Correlation coefficient indicates the extent to which two variables move together. Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
- Correlation have objective to find a numerical value expressing the relationship between variables. Regression have objective to estimate values of random variable on the basis of the values of fixed variable.

Coefficient of Determination

The ratio of the unexplained variation to the total variation represents the proportion of variation in Y that is not explained by regression on X. Subtraction of this proportion from 1.0

gives the proportion of variation in Y that is explained by regression on X. The statistic used to express this proportion is called the coefficient of determination and is denoted by R^2. It may be written as follows :

Focus Formula

$$R^2 = 1 - \frac{\text{Variation in Y remaining after regression on X}}{\text{Total variation in Y}}$$

$$R^2 = 1 - \frac{\text{Error sum of squares}}{\text{Total sum of squares}}$$

The value of R^2 is the proportion of the variation in the dependent variable Y explained by regression on the independent variable X.

Example : After investigation it has been found the demand for automobiles in a city depends mainly, if not entirely, upon the number of families residing in that city. Below are given figures for the sales of automobiles in the five cities for the year 2003 and the number of families residing in those cities.

City	No. of Familiesin Lakhs (X)	Sale of Automobilesin 000's (Y)
A	70	25.2
B	75	28.6
C	80	30.2
D	60	22.3
E	90	35.4

Fit a linear regression equation of Y on X by the least square method and estimate the sales for the year 2006 for city A which is estimated to have 100 lakh families assuming that the same relationship holds true.

Solution : Calculation of Regression Equation

City	X	Y	X^2	XY
A	70	25.2	4,900	1,764
B	75	28.6	5,625	2,145
C	80	30.2	6,400	2,416
D	60	22.3	3,600	1,338
E	90	35.4	8,100	3,186
	$\Sigma X = 375$	$\Sigma Y = 141.7$	$\Sigma X^2 = 28,625$	$\Sigma XY = 10,849$

Regression equation of Y on X is

$$Y = a + bX.$$

To determine the values of a and b, we shall solve the normal equations

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values from the table, the normal equations become

$141.7 = 5a + 375b$ ...(i)

$10,849 = 375a + 28,625b$ ...(ii)

Multiplying Eqn. (i) by 75 and subtracting form Eqn. (ii). we get

$221.5 = 500b$ or $b = 0.443$

Substituting the value of b in Eqn. (i), we have

$-24.425 = 5a$ or $a = -4.885$

Therefore, the regression equation of Y on X is

$y = -4.885 + 0.443X$

Estimated sales for the year 2006 for city A

$Y = -4.885 + 0.443\ (100)$

$= -4.885 + 44.3 = 39.415$

Hence it is expected that about 39.415 autos would be sold in city A having a population of 100 lakh families.

Example : If regression coefficient of x on y = $-\frac{1}{6}$ and that of y on x = $-\frac{3}{2}$. Find the value of correlation coefficient between x and y series.

Solution :

$\because \quad r = \sqrt{(b_{xy})(b_{yx})}$

or $\quad r = \sqrt{\frac{1}{6} \times \frac{3}{2}}$

or $\quad r = \sqrt{-.0167 \times -1.5} = 0.15$

since b_{yx} and b_{xy} both are negative, so coefficient of correlation is also negative.

Example : In a correlation study the following values are obtained :

	X	Y
Mean	65	67
S.D.	2.5	3.5

Coefficient of Correlation r = 0.8

Find the two regression equations.

Solution : Regression equation of X on Y :

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$\bar{X} = 65$, $\sigma_x = 2.5$, $\sigma_y = 3.5$, $r = 0.8$. $\bar{Y} = 67$

$$X - 65 = 0.8\frac{2.5}{35}(Y - 67)$$

$X - 65 = 0.571\ (Y - 67)$

$X - 65 = 0.571\ Y - 38.26$

$X = 0.571\ Y + 26.74$

Regression equation of Y on X :

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$Y - 67 = 0.8\frac{3.5}{2.5}(X - 65)$$

$Y - 67 = 1.12\ (X - 65)$

$Y - 67 = 1.12\ X - 72.8$

$Y = 1.12\ X - 5.8$